```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.43 on 594 degrees of freedom
Multiple R-squared:  0.2155,Adjusted R-squared:  0.2103
F-statistic:  40.8 on 4 and 594 DF,  p-value: < 2.2e-16
```

De variabele `age` heeft de grootste p-waarde (0.0938) en deze is groter dan 0.01. We voeren dezelfde analyse zonder `age`.

```
> summary(lm(wt ~ gestation+dwt+number, data=geboorte))

Call:
lm(formula = wt ~ gestation + dwt + number, data = geboorte)

Residuals:
    Min      1Q  Median      3Q     Max 
-50.213 -10.439  -0.934   9.805  55.574 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -26.23421   12.78567  -2.052   0.0406 *  
gestation     0.45218    0.04223  10.708  < 2e-16 ***
dwt           0.12734    0.02951   4.316 1.86e-05 ***
number       -0.29694    0.06198  -4.791 2.10e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 16.46 on 595 degrees of freedom
Multiple R-squared:  0.2118,Adjusted R-squared:  0.2079
F-statistic:  53.3 on 3 and 595 DF,  p-value: < 2.2e-16
```

Alle variabelen hebben een p-waarde kleiner dan 0.01 en de achterwaartse selectie stopt hier. De geselecteerde predictoren zijn `gestation`, `dwt` en `number`.

10.2 Lineaire regressie met nominale predictoren

Stel dat we een model willen analyseren waarbij een predictor nominaal is, met meer dan twee niveaus. We gaan die nominale predictor vervangen door meerdere 0-1 predictoren, die we hulpveranderlijken gaan noemen. In het algemeen geldt dat we een nominale predictor met I niveaus moeten hercoderen tot $I-1$ nieuwe hulpveranderlijken (0-1 variabelen) die we vervolgens in het regressiemodel kunnen stoppen. We illustreren dit met een nieuw voorbeeld.

Reactietijd bij bepaalde cognitieve taken spelen een belangrijke rol bij het begrijpen van depressie (zie bv. Kaiser et al. [2008]). Je bent onderzoeker in klinische psychologie en je wil vier types behandeling vergelijken en je gebruikt een steekproef van 37 patiënten. Ze worden in vier gerandomiseerde groepen ingedeeld en ze volgen één van de vier behandelingen A, B, C of D. Na drie maanden worden ze uitgenodigd om een aantal cognitieve taken uit te voeren. In het data frame `depressie` vind je de reactietijden van de 37 patiënten bij één van die cognitieve taken.

```
> depressie
   behandeling reactietijd
1            A       0.925
2            D       0.875
3            A       0.825
4            B       0.950
...        ...         ...
36           C       1.170
37           C       1.155
```

Daar de variabele `behandeling` nominaal is met meer dan twee niveau's mogen we niet zomaar een lineaire regressie uitvoeren om te weten of verschillen in `reactietijd` verklaard kunnen worden door `behandeling`. We kunnen ook geen t-toets van hoofdstuk 6 gebruiken omdat de t-toets om verwachtingen te vergelijken alleen met twee groepen werkt, niet met meer dan twee. We gaan dus de variabele `behandeling` hercoderen tot 3 ($= 4 - 1$) nieuwe hulpveranderlijken met twee niveaus: 0 en 1.

10.2.1 Hercodering

We kunnen een onderscheid maken tussen twee hercoderingen: dummy-codering en effect-codering.

- Bij dummy-codering kiest men 1 van de I niveaus als referentieniveau en worden de andere niveaus via een 0-1 variabele gecodeerd.

  In het geval van het voorbeeld betekent dit dat we 3 hulpveranderlijken X_1, X_2 en X_3 moeten aanmaken. Wanneer we behandeling D als referentieniveau beschouwen[3], dan bekomen we de volgende codering:

Behandeling	X_1	X_2	X_3
A	1	0	0
B	0	1	0
C	0	0	1
D	0	0	0

  Dit betekent concreet dat voor een individu i

  - die behandeling A volgt, geldt: $x_{i1} = 1$, $x_{i2} = 0$, $x_{i3} = 0$.
  - die behandeling B volgt, geldt: $x_{i1} = 0$, $x_{i2} = 1$, $x_{i3} = 0$.

[3] De keuze van het referentieniveau is vrij

- die behandeling C volgt, geldt: $x_{i1} = 0$, $x_{i2} = 0$, $x_{i3} = 1$.
- die behandeling D volgt, geldt: $x_{i1} = 0$, $x_{i2} = 0$, $x_{i3} = 0$.

- Bij effect-codering wordt ook een groep gekozen maar deze groep wordt steeds met -1 gecodeerd i.p.v. met 0. Deze groep wordt niet als referentie beschouwd. Voor het voorbeeld bekomen we:

Behandeling	X_1	X_2	X_3
A	1	0	0
B	0	1	0
C	0	0	1
D	-1	-1	-1

  Dit betekent dat de codering hetzelfde is als de dummy-codering voor individuen die behandeling A, B of C volgen maar voor een individu i die behandeling D volgt, geldt: $x_{i1} = -1$, $x_{i2} = -1$, $x_{i3} = -1$.

Het effect van de behandeling op de verwachting van de reactietijd Y kunnen we als volgt modelleren:

$$E(Y_i \,|\, x_{i1}, x_{i2}, x_{i3}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}.$$

Naargelang het coderingsschema dat gehanteerd wordt, hebben de regressieparameters een andere betekenis.

- Dummy-codering
  - $E(Y_i \,|\, \mathrm{D}) = E(Y_i \,|\, x_{i1} = 0, x_{i2} = 0, x_{i3} = 0) = \beta_0$. De coëfficiënt β_0 stelt dus de verwachte reactietijd voor bij behandeling D.
  - $E(Y_i \,|\, \mathrm{A}) = E(Y_i \,|\, x_{i1} = 1, x_{i2} = 0, x_{i3} = 0) = \beta_0 + \beta_1$. De coëfficiënt β_1 stelt dus het verschil voor van de verwachte reactietijd bij behandeling A en de verwachte reactietijd bij behandeling D.
  - Analoog stellen β_2 en β_3 het verschil in verwachte reactietijd tussen behandeling B en behandeling D en tussen behandeling C en behandeling D.

- Effect-codering
  - In dit geval kan aangetoond worden dat β_0 het *marginale gemiddelde*[4] van de reactietijd voorstelt, i.e. het gemiddelde van de verwachtte reactietijden over de verschillende behandelingen heen:

$$\beta_0 = \big(\ E(Y_i \,|\, \mathrm{A}) + E(Y_i \,|\, \mathrm{B}) + E(Y_i \,|\, \mathrm{C}) + E(Y_i \,|\, \mathrm{D})\ \big)/4.$$

110. Bereken het marginale gemiddelde van `reactietijd`*.*

[4]Let op; dezelfde uitdrukking “marginale gemiddelde” wordt gebruikt voor het gemiddelde van de verwachtingen en voor het gemiddelde van de corresponderende gemiddelden.

– $E(Y_i \,|\, \mathrm{A}) = E(Y_i \,|\, x_{i1} = 1, x_{i2} = 0, x_{i3} = 0) = \beta_0 + \beta_1$. De coëfficiënt β_1 stelt dus het verschil voor tussen de verwachte reactietijd bij behandeling A en het marginale gemiddelde.

– In het algemeen, β_ℓ ($\ell = 1, 2, 3$) drukt het verschil uit tussen de verwachte reactietijd bij behandeling ℓ en het marginale gemiddelde.

– De verwachte reactietijd bij behandeling D is

$$E(Y_i \,|\, \mathrm{D}) = E(Y_i \,|\, x_{i1} = -1, x_{i2} = -1, x_{i3} = -1) = \beta_0 - \beta_1 - \beta_2 - \beta_3.$$

Dit betekent dat het verschil tussen de verwachte reactietijd bij behandeling D en het marginale gemiddelde gelijk is aan $-\beta_1 - \beta_2 - \beta_3$.

10.2.2 Welke hypothese?

De variabele X_1 heeft geen betekenis in zich. Stel dat we weten dat $x_{1i} = 0$ bij individu i. Dit geeft ons geen duidelijke informatie over dat individu. Hetzelfde geldt voor elke hulpveranderlijke. We gaan dus nooit toetsen of β_j al dan niet nul is. Als we modellen vergelijken, gaan we ook nooit een model beschouwen met de hulpveranderlijke X_1 en zonder de hulpveranderlijke X_2. Idem bij predictorenselectie: we beschouwen alleen modellen die alle hulpveranderlijken bevatten of geen.

Dus, indien we willen toetsen of een nominale variabele een predictor van Y is, dan gaan we het model met alle hulpveranderlijken vergelijken met het model zonder de hulpveranderlijken, a.d.h.v. een F-toets. Merk op dat het resultaat van deze toets onafhankelijk is van het gehanteerde coderingsschema: dummy-codering en effect-codering geven dezelfde resultaten. Het resultaat is ook onafhankelijk van het gekozen referentieniveau.

10.2.3 Berekeningen met `R`

Dankzij `R` hoeven we niet zelf hulpveranderlijken te definiëren. We hoeven ook niet zelf een coderingsschema en een referentieniveau te kiezen. `R` doet het allemaal voor ons. We gaan nog de functie `lm` gebruiken en als één (of meerdere) van de predictoren nominaal is, gaat `R` automatisch dummy-codering gebruiken met het eerste niveau als referentie. `R` gaat ook zelf hulpveranderlijken definiëren. Voorbeeld:

```
> LM.depressie <- lm( reactietijd ~ behandeling, data = depressie)
> summary(LM.depressie)

Call:
lm(formula = reactietijd ~ behandeling, data = depressie)

Residuals:
     Min       1Q   Median       3Q      Max
-0.15812 -0.04550 -0.01611  0.05889  0.13050
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.91111    0.02744  33.207  < 2e-16 ***
behandelingB  0.07839    0.03782   2.073  0.04608 *
behandelingC  0.27939    0.03782   7.387 1.74e-08 ***
behandelingD  0.12201    0.04000   3.051  0.00448 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.08231 on 33 degrees of freedom
Multiple R-squared:  0.6418,Adjusted R-squared:  0.6093
F-statistic: 19.71 on 3 and 33 DF,  p-value: 1.674e-07
```

111. Ga de normaliteitsassumptie, de 1ste en de 2de Gauss-Markov assumpties na.

Het commando is vanzelfsprekend. De eerste regels van de output zijn zoals vroeger. De tabel met de coëfficiënten bevat vier regels: één per hulpveranderlijke. In de regel van het intercept vinden we $\hat{\beta}_0 = 0.91111$: de schatting van de verwachte reactietijd bij behandeling A. De corresponderende p-waarde is, zoals bijna altijd, niet relevant. In de regel `behandelingB` vinden we $\hat{\beta}_B = 0.07839$: de schatting van het verschil tussen de verwachte reactietijd bij behandelingen A en B. De corresponderende p-waarde is niet betekenisvol omdat één hulpveranderlijke in zich geen betekenis heeft. Dan hebben we nog twee analoge regels voor behandelingen B en C. Helemaal onderaan vinden we de p-waarde van de F-toets die ons model vergelijkt met het nulmodel. Deze p-waarde (`1.674e-07`) is wel relevant. Ze is kleiner dan 0.05 en we besluiten dus dat `behandeling` een predictor is van `reactietijd`. Daar de p-waarde veel kleiner is dan 0.05 is de schending van homoscedasticiteit (zie oefening 111) niet belangrijk.

112. Gebruikmakend van dezelfde codering als `R`, wat zijn de waarden van de hulpveranderlijken X_1, X_2 en X_3 bij individu 27?

Er zijn contexten waar effect-codering handiger is dan dummy-codering. Het is dan mogelijk om dat coderingschema te hanteren bij de berekeningen met `R`; Dit wordt in deze cursus niet gezien.

10.2.4 Een voorbeeld met meerdere predictoren — `sportData`

We willen nagaan of de variabele `type` een predictor van `tijd` is, rekening houdend met `lengte` en `sport`. De variabele `type` is een nominale variabele met vijf niveaus: `andere`, `basketbal`, `tennis`, `voetbal` en `zwemmen`. Het referentieniveau zal dus `andere` zijn en `R` gaat vier hulpveranderlijken definiëren. Laten we het model met de drie predictoren analyseren.

```
> LM.sport <- lm(tijd ~ lengte + sport + type, data = sportData)
> summary(LM.sport)

Call:
lm(formula = tijd ~ lengte + sport + type, data = sportData)

Residuals:
    Min      1Q  Median      3Q     Max
```

```
-9.8812 -2.8280 -0.2145  2.6244  9.7401

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   24.97628    3.47839   7.180 1.47e-11 ***
lengte        -0.04436    0.01887  -2.350   0.0198 *
sport          1.75088    0.28944   6.049 7.42e-09 ***
typebasketbal  0.82205    1.04452   0.787   0.4322
typetennis     0.83061    1.00160   0.829   0.4080
typevoetbal    0.86430    0.88839   0.973   0.3318
typezwemmen   -0.17403    0.99309  -0.175   0.8611
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.268 on 193 degrees of freedom
Multiple R-squared:  0.1987,Adjusted R-squared:  0.1737
F-statistic: 7.975 on 6 and 193 DF,  p-value: 1.076e-07
```

Daar er meerdere predictoren zijn, kunnen we hier het intercept (24.97628) niet interpreteren als de schatting van $\mu_{\texttt{tijd}}$ bij individuen die aan een andere sport doen. De correcte interpretatie is: het intercept (24.97628) is de schatting van $\mu_{\texttt{tijd}}$ bij individuen die aan een andere sport doen <u>en</u> waarbij de variabelen `lengte` en `sport` nul zijn. Zulke individuen bestaan uiteraard niet en het intercept heeft dus geen concrete en intuïtieve betekenis.

De coëfficiënt van `lengte` is -0.04436. Dit is de schatting van $\beta_{\texttt{lengte}}$. Dit betekent dat twee individuen met een verschil van één cm op `lengte` en met identieke scores op alle andere variabelen een verschil van -0.04436 seconde zullen ervaren op `tijd` (gemiddeld gezien). Daar de coëfficiënt negatief is, zal `tijd` lager zijn bij het langere individu. Het langere individu loopt dus sneller (gemiddeld gezien).

De coëfficiënt van `typebasketbal` (0.82205) representeert het gemiddelde tijdsverschil tussen een individu die aan basketbal doet en een individu die aan "andere" doet, indien ze identieke scores hebben op de andere variabelen. Een analoge interpretatie geldt voor de coëfficiënten van `typetennis`, `typevoetbal` en `typezwemmen`.

De p-waarde (`1.076e-07`) van de F-toets helemaal onderaan de output heeft niets te maken met onze onderzoeksvraag (is `type` een predictor van `tijd`, rekening houdend met `lengte` en `sport`?). Deze p-waarde heeft betrekking tot de vergelijking van het nulmodel (zonder predictor) met het model met drie predictoren.

Om onze onderzoeksvraag te beantwoorden moeten we het model met drie predictoren vergelijken met hetzelfde model maar zonder `type` We maken dus nu een lineair model aan zonder de predictor `type`, maar wel met de twee andere predictoren..

```
> LM.sportZonderType <- lm(tijd ~ lengte + sport, data = sportData)
```

Nu kunnen we beide modellen vergelijken m.b.v. een F-toets, dankzij de functie `anova`.

```
> anova(LM.sportZonderType,LM.sport)
Analysis of Variance Table

Model 1: tijd ~ lengte + sport
Model 2: tijd ~ lengte + sport + type
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1    197 3556.1
2    193 3514.9  4    41.155 0.5649 0.6884
```

De p-waarde (0.6884) van deze toets is groter dan 0.05 en we besluiten dat `type` geen predictor van `tijd` is.

10.2.5 Nog een voorbeeld — `microbusiness`

In de Verenigde Staten zijn er veel programma's om vrouwen met een laag inkomen die een microbusiness stichten financieel te ondersteunen. In een onderzoek [Sanders, 2004] wil de auteur nagaan of die programma's efficiënt zijn. Drie steekproeven worden getrokken: een steekproef van vrouwen die de steun van zo'n programma krijgen ($n = 62$), een steekproef van vrouwen die een microbusiness hebben maar geen steun krijgen ($n = 57$) en een steekproef van vrouwen die geen microbusiness hebben maar wel werken ($n = 178$). De toename (of afname) van het inkomen tussen 1991 en 1995 wordt geregistreerd, samen met het ras. De gegevens (data frame `microbusiness`) zien er als volgt uit:

```
> microbusiness
            groep inkomenWijziging   race
1          GeenMB             8941 latino
2          GeenMB            -5798  black
3          GeenMB            19240  black
4          GeenMB            -9746  white
5          GeenMB             3023  black
6      MBMetSteun             8200  black
...           ...              ...    ...
295 MBZonderSteun            19712  white
296        GeenMB            22082  white
297        GeenMB             2656 latino
```

We gaan eerst de gemiddelde inkomenwijziging in de drie groepen berekenen. Als we het commando `mean(microbusiness$inkomenWijziging)` gebruiken, dan komen we het gemiddelde van alle vrouwen. Dat is niet wat we nodig hebben. Om de gemiddelden in de drie groepen apart te krijgen gebruiken we de functie `aggregate`:

```
> aggregate( formula = microbusiness$inkomenWijziging ~ microbusiness$groep,
```

```
   FUN = mean )
  microbusiness$groep microbusiness$inkomenWijziging
1              GeenMB                            8652
2          MBMetSteun                            5708
3       MBZonderSteun                            6455
```

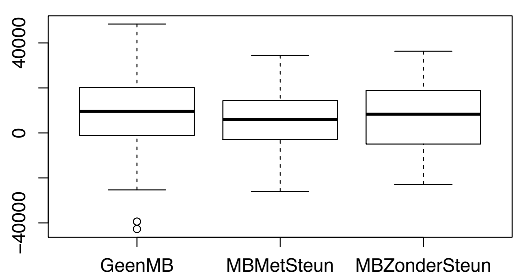
Het argument `formula` werkt zoals in het vorige hoofdstuk. Het maakt duidelijk dat we wensen de variabele `inkomenWijziging` te voorspellen m.b.v. de predictor `groep`. Het argument `FUN` is de afkorting van 'functie' en wordt gebruikt om `R` te zeggen wat hij moet berekenen in elke groep. Met het argument `FUN = mean` weet `R` dat hij het gemiddelde in elke groep moet berekenen. Met het argument `FUN = var` zou `R` de schatting van de variantie in elke groep berekenen. Met `FUN = median` zou `R` de mediaan berekenen. Enz.

113. Voer het commando `aggregate( tijd geslacht+type, FUN = mean, data = sportData)` *uit. Begrijp je de output?*

We kijken nu naar de output van het commando. We zien dat de drie gemiddelden niet identiek zijn en dat de verschillen niet gering zijn (ongeveer 3000$ tussen groep 1 en 2). Kunnen we hieruit afleiden dat de drie verwachtingen in de populaties niet identiek zijn? Niet zomaar. Laten we voorzichtig zijn en laten we de gegevens visueel analyseren met de `boxplot` functie:

```
> boxplot(formula=microbusiness$inkomenWijziging ~ microbusiness$groep)
```

De output wordt in Fig. 10.1 weergegeven. De boxplots tonen dat de medianen ook van elkaar verschillen (ongeveer zoals de gemiddelden) maar vooral dat



Figuur 10.1: Boxplot van de inkomenwijzigingen in de drie groepen.

de variatie binnen elke steekproef zeer groot is: veel groter dan de verschillen tussen de medianen of tussen de gemiddelden. De verschillen tussen de medianen lijken bijna verwaarloosbaar t.o.v. de variatie binnen elke steekproef. Dit geeft de indruk dat de verschillen tussen de drie groepen toevallig zijn.

114. Gebruik de functies `aggregate` *en de* `boxplot` *om de verdeling van de toevalsvariabele* `score` *te vergelijken in de drie opleidingen psy, ped en soc, m.b.v. het data frame* `myData`.

We berekenen nu de gemiddelden bij de drie rassen.

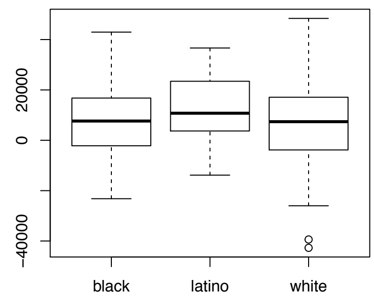
```
> aggregate( formula=inkomenWijziging ~ race,FUN = mean,
  data=microbusiness)
    race inkomenWijziging
1  black         7460.496
2 latino        12168.581
3  white         6709.917
```

Er zijn grote verschillen tussen de drie rassen. Laten we nu de boxplots tekenen.

```
boxplot(formula=microbusiness$inkomenWijziging ~ microbusiness$race)
```

115. Maak een lineair model aan om `score` *te verklaren m.b.v.* `opleiding`. *Gebruik de functie* `plot` *om de homoscedasticiteitsassumptie na te gaan. Vergelijk met de boxplot van oefening 114.*

De output wordt in Fig. 10.2 weergegeven. Zoals bij Fig. 10.1 lijken de verschillen tussen de medianen bijna verwaarloosbaar t.o.v. de variatie binnen elke steekproef. Dit geeft de indruk dat de verschillen tussen de drie rassen ook toevallig zijn. We gaan een lineair model gebruiken om de verschillen tussen



Figuur 10.2: Boxplot van de inkomenwijzigingen bij de drie rassen.

groepen en rassen te verklaren en we gaan dit model toetsen. We willen dus een lineair model toetsen met `groep` en `race` als predictoren. Omdat beide predictoren nominaal zijn, worden ze ook factor genoemd.

```
> LM.mb <- lm(inkomenWijziging ~ groep + race, data = microbusiness)
> summary(LM.mb)

Call:
lm(formula = inkomenWijziging ~ groep + race, data = microbusiness)

Residuals:
   Min     1Q Median     3Q    Max
-50418  -9537    412  10428  40684

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)          8486.6     1503.2   5.646  3.9e-08 ***
groepMBMetSteun     -2773.1     2227.3  -1.245    0.214
groepMBZonderSteun  -2200.2     2298.4  -0.957    0.339
racelatino           4555.0     3011.4   1.513    0.131
racewhite            -770.7     1852.5  -0.416    0.678
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15090 on 292 degrees of freedom
Multiple R-squared:  0.01786,Adjusted R-squared:  0.004411
```

```
F-statistic: 1.328 on 4 and 292 DF,  p-value: 0.2596
```

Zoals bij de andere voorbeelden met nominale variabelen heeft `R` hulpveranderlijken gedefinieerd: twee om `groep` te hercoderen en twee om `race` te hercoderen. Het referentieniveau is alfabetisch bepaald: `GeenMB` voor de groep en `black` voor het ras.

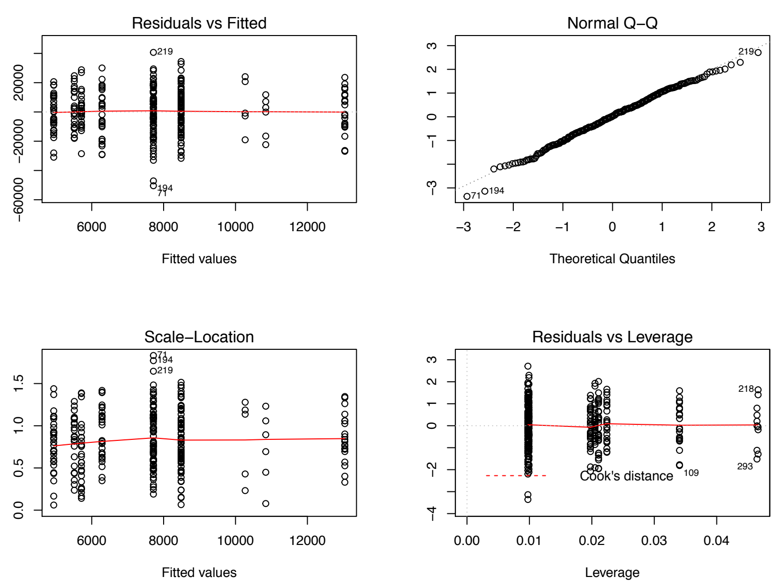
Laten we de output van `summary(LM.mb)` bespreken. In de rij van het intercept lezen we 8486.6 af. Dit is $\hat{\beta}_0$: de schatting van de voorwaardelijke verwachting van Y voor zwarte vrouwen die geen microbusiness hebben.

Wat is de schatting van de voorwaardelijke verwachting van Y voor zwarte vrouwen die wel een microbusiness hebben, maar geen steun? Het is $8486.6 - 2200.2 = 6286.4$. En de schatting van de voorwaardelijke verwachting van Y voor latino vrouwen die een microbusiness hebben, zonder steun? Het is $8486.6 - 2200.2 + 4555.0 = 10841.4$. Enz.

De p-waarde van de F-toets is 0.2596. Ons lineair model maakt dus predicties die niet significant beter zijn dan die van het nulmodel en we aanvaarden dus het nulmodel.

We kijken nu naar de output (Fig. 10.3) van het commando `plot(LM.mb)` om de assumpties van lineaire regressie na te gaan.

```
> plot(LM.mb)
Hit <Return> to see next plot:
```



Figuur 10.3: Assumptiecheck: output van `plot(LM.mb)`.

De "`Residuals vs Fitted`" plot ziet er niet uit zoals bij de andere voorbeelden. We zien niet echt een puntenwolk, maar verticale lijnen. De reden is dat de predictoren nominaal zijn. Het lineair model maakt dan een beperkt aantal predicties: één per combinatie van de niveaus van de factoren. We hebben bij dit voorbeeld twee factoren, elk met drie niveaus: dus $3 \times 3 = 9$ combinaties en er zijn inderdaad 9 verticale lijnen. Voor de rest is de interpretatie van deze grafiek zoals vroeger. De rode curve is min of meer horizontaal en de eerste Gauss-Markov assumptie is dus in orde.

De normale qq-plot is in orde.

Op de "`Scale-Location`" plot zien we ook 9 verticale lijnen omdat de predictoren nominaal zijn. Voor de rest is de rode curve min of meer horizontaal en de tweede Gauss-Markov assumptie is dus in orde. De vierde grafiek wordt niet besproken.

10.3 Historische nota — variantie-analyse (anova)

Een techniek om verwachtingen in twee groepen te vergelijken werd in het begin van de 20ste eeuw ontwikkeld: de t-toets (zie Rubr. 6.5.2). Deze techniek werd in de eerste helft van de 20ste eeuw veralgemeend om verwachtingen in p groepen te vergelijken. Deze techniek heet variantie-analyse (analysis of variance, anova). Indien de groepen bepaald worden op basis van één nominale variabele, dan spreekt men van one-way anova. Bv. zijn de verwachte scores op het examen Statistiek II identiek in de groepen van psychologie studenten, ped. wetenschappen studenten en sociaal werk studenten? De nominale variabele van belang is `opleiding`.

Indien de groepen bepaald worden op basis van twee nominale variabelen, dan spreekt men van two-way anova. Bv. zijn de verwachte scores op het examen Statistiek II identiek in de groepen van vr. psy., man. psy., vr. ped., man. ped., vr. soc. en man. soc. studenten? De nominale variabelen van belang zijn nu `opleiding` en `geslacht`. Als we die twee nominale variabelen (of factoren) kruizen, dan komen we 6 ($= 3 \times 2$) groepen uit. Three-way, four-way, ... anova worden op dezelfde manier gedefinieerd.

Bij een variantie-analyse wordt, zoals bij een t-toets, nagegaan of de verschillen tussen de groepen het effect van het toeval kunnen zijn (nulhypothese) of niet. Schattingen van de verwachtingen in elke groep worden dan berekend en ze kunnen gebruikt worden om predicties te maken. Bv. als student A een man is en psychologie studeert, dan kan je voorspellen dat zijn score op Statistiek II gelijk zal zijn aan de geschatte verwachting van de corresponderende groep. Met regressie-analyse met nominale variabelen doen we eigenlijk hetzelfde en het is mogelijk te bewijzen dat beide technieken equivalent zijn, met nominale variabelen. De p-waarde van de F-toets bij een variantie-analyse is identiek aan de p-waarde van de F-toets bij een lineaire regressie. Maar lineaire regressie laat ook toe om continue predictoren van ratio of interval meetniveau te gebruiken. Lineaire regressie is dus algemener (of krachtiger) en er is geen reden om beide technieken te studeren en te gebruiken. In deze cursus wordt dus geopteerd om geen variantie-analyse te zien.

Daar de berekeningen simpler zijn bij variantie-analyse dan bij lineaire re-